Valley: Video Assistant with Large Language model Enhanced abilitY

Year: 2023 | *Citations: 246* | *Authors: Ruipu Luo, Ziwang Zhao, Min Yang*

Abstract

Large Language Models (LLMs), with remarkable conversational capability, have emerged as AI assistants that can handle both visual and textual modalities. However, their effectiveness in joint video and language understanding has not been extensively explored. In the paper, we introduce Valley, a multi-modal foundation model that is designed to enable enhanced video comprehension and instruction-following capabilities. To this end, we construct two datasets, namely Valley-702k and Valley-instruct-73k, to cover a diverse range of video-text alignment and video-based instruction tasks, such as multi-shot captions, long video descriptions, action recognition, causal inference, etc. Then, we adopt ViT-L/14 as the vision encoder and explore three different temporal modeling modules to learn multifaceted features for enhanced video understanding. In addition, we implement a two-phase training approach for Valley: the first phase focuses solely on training the projection module to facilitate the LLM's capacity to understand visual input, and the second phase jointly trains the projection module and the LLM to improve their instruction following ability. Extensive experiments demonstrate that Valley has the potential to serve as an effective video assistant, simplifying complex video-understanding scenarios. Our code and data are published anonymously at https://github.com/valley-vl/Valley.